Master Project: Learnable Tokenization for Deep Vision Models

Torstein Forseth Marius Aasan

December 2023

1 Overview

In this project, we look to explore learnable tokenization as a preprocessing step for vision models, with particular focus on Vision Transformers (ViTs) [Vaswani et al., 2017, Dosovitskiy et al., 2021] and Graph Neural Networks (GNNs) [Kipf and Welling, 2017, Veličković et al., 2018, Xu et al., 2019].

Building on the work in SPiT: Superpixel Transformers [Anonymous, 2023], we are fundamentally interested in exploring differentiable methods for superpixel tokenization that can be leveraged for improved accuracy and efficiency in modelling tasks.

2 Differentiable Tokenization

In this section, we begin by outlining the theoretic framework for superpixels, and their role in tokenization for ViTs. We look at some existing methods and potential baselines, while proposing some potential promising research directions for the study.

2.1 Superpixels and Partitions

We begin by letting $H \times W = \big((y, x) : 1 \leq y \leq h, 1 \leq x \leq w\big)$ denote the coordinates of an image of spatial dimension (h, w). For notational convenience, we let $\mathcal{I}$ be an index set for the mapping $i \mapsto (y, x)$. This allows us to consider a C-channel image as a signal $\xi : \mathcal{I} \to \mathbb{R}^C$.

Let $E^{(0)} \subset \mathcal{I} \times \mathcal{I}$ denote the edges for four-way adjacency under $H \times W$. We consider a superpixel P as a set $P \subset \mathcal{I}$, and we say that P is connected if for any two pixels $p, q \in P$, there exists a sequence of edges in $((i_j, i'_j) \in E^{(0)})_{j=1}^{k}$ such that $i_1 = p$ and $i'_k = q$. A set of superpixels form a partition π of an image if for any two distinct superpixels $P, P' \in \pi$, their intersection $P \cap P' = \emptyset$, and the union of all superpixels is equal to the set of all pixel positions in the image, i.e., $\bigcup_{P \in \pi^{(t)}} P = \mathcal{I}$.

2.2 Differentiable Tokenization

Vision Transformers canonically rely on partitioning and tokenization of an image for modelling. This process was originally proposed to consist of partitioning the image into square patches of a fixed size, followed by vectorization of patches using standard scanning order, before projecting the patch features via a learnable embedding. In the SPiT model [Anonymous, 2023], the authors outline a method for generalizing this partitioning beyond square patches, and propose an effective heuristic on-line tokenization scheme. This is notably a non-learnable non-differentiable operation, and this work looks to expand on this.

There are several paths for potential exploration to expand the method to differentiable tokenization. We begin by outlining a few methods that have recently been proposed.

- One existing baseline comes from differentiable SLIC [Zhu et al., 2023] which was successfully applied to semantic segmentation tasks.
- ToMe [Bolya et al., 2023] proposes a token merging procedure in the transformer blocks to combine tokens for efficiency.
- MSViT [Havtorn et al., 2023] proposes a learnable function between two levels of patch granularity, for similar reasons.
- Quadformer [Ronen et al., 2023] and Quadtreenet [Chitta et al., 2020] are methods that look to exploit quadtree representations in tokenization and partitioning for vision models.

Clearly, the recent interest in dynamic tokenization for transformers provides ample opportunities for further investigations, and provides baselines for contrasting approaches in our study. However, we also look to generalized superpixel methods which could be extendable to a learnable framework, including

- One approach to learnable tokenization is to simply extend SPiT to a GNN framework. This has the additional benefit of providing a hierarchical partitioning.
- Another approach to learnable tokenization is to leverage an energy potential between pixels using Ising models. Learning optimal partition functions can then be done via line graphs or max-cut algorithms for binary partitioning, or extended to a general Potts model with multi-state phase transition.
- ETPS Liu et al. [2011] applies a MRF/CRF with an energy function, which can be amenable to differentiable learning.
- Other graph based approaches have been proposed, including POISE [Humayun et al., 2015]. This method expands on min-cut methods for producing partitions.

3 Literature, Work Packages, and Timeline

We subdivide the work into work packages, each with a different focus, and loosely building on the work of the previous task. This helps ensure the project's successful completion, and provides checkpoints for the progression of the project.

3.1 Literature Search

For kickstarting the literature search, we outline the following readings in addition to the literature outlined in the previous section.

- Superpixels - Evaluation of State of the Art, Stutz et al. [2018];
- Digital Geometry, Klette and Rosenfeld [2004];
- Hierarchical Graph Representation Learning with Differentiable Pooling, Ying et al. [2019].

The candidate will also be encouraged to do independent literature search, particularly during work packages (A)-(B), as outlined in the subsequent section.

3.2 Work Packages

We subdivide the main thesis work into three work packages.

- Package (A) involves carrying out initial experimental baselines, implementation of various heuristic methods, and evaulation of potential applicablity to learnable tokenization and partitioning. We decide on suitable datasets for the task, and look at various image modalities, and begin writing on the theoretical underpinnings of the project.
- Package (B) is concerned with carrying out more extensive evaluation of different methods, more literature search, and establishing one or more promising methods for the final evaluation. Here, the candidate will be able to showcase their ability for independent research, and critical thinking in narrowing down potential directions.
- Package (C) will revolve around deciding on a well-rounded set of tests and ablations to evaluate the proposed solution in context of existing works. It also involves potentially expanding proposed methods to self-supervised or unsupervised learning paradigms.

Semester	Activity
Fall 2023	
Fall 2023	
Fall 2023	
Spring 2024	
Spring 2024	
Spring 2024	Literature study
Spring 2024	Start work on module A
Fall 2024	
Fall 2024	Finalize module A
Fall 2024	Start work on module B
Spring 2025	Finalize module B
Spring 2025	Finalize module C

Table 1: Timeline for Courses and Work Packages

References

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Adv. Neural Inf. Process. Sys. (NeurIPS)*, pages 5998–6008, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Inter. Conf. Learn. Represent. (ICLR)*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Inter. Conf. Learn. Represent. (ICLR)*, 2017. URL `https://openreview.net/forum?id=SJU4ayYgl`.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *Inter. Conf. Learn. Represent. (ICLR)*, 2018. URL `https://openreview.net/forum?id=rJXMpikCZ`.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *Inter. Conf. Learn. Represent. (ICLR)*, 2019. URL `https://openreview.net/forum?id=ryGs6iA5Km`.

Anonymous. A spitting image: Superpixel transformers. In *Submitted to The Twelfth International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=Vy6sjPt2Vr`. under review.

Alex Zihao Zhu, Jieru Mei, Siyuan Qiao, Hang Yan, Yukun Zhu, Liang-Chieh Chen, and Henrik Kretzschmar. Superpixel transformers for efficient semantic segmentation, 2023.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *Inter. Conf. Learn. Represent. (ICLR)*, 2023. URL https://openreview.net/forum?id=JroZRaRw7Eu.

Jakob Drachmann Havtorn, Amélie Royer, Tijmen Blankevoort, and Babak Ehteshami Bejnordi. Msvit: Dynamic mixed-scale tokenization for vision transformers. In *IEEE Inter. Conf. Comput. Vis. (ICCV)*, pages 838–848, October 2023.

Tomer Ronen, Omer Levy, and Avram Golbert. Vision transformers with mixed-resolution tokenization. In *IEEE/CVF Inter. Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 4612–4621, 2023.

Kashyap Chitta, José M. Álvarez, and Martial Hebert. Quadtree generating networks: Efficient hierarchical scene parsing with sparse convolutions. In *IEEE Wint. Conf. Appl. Comput. Vis. (WACV)*, pages 2009–2018. IEEE, 2020. doi: 10.1109/WACV45572.2020.9093449. URL https://doi.org/10.1109/WACV45572.2020.9093449.

Ming-Yu Liu, Oncel Tuzel, Srikumar Ramalingam, and Rama Chellappa. Entropy rate superpixel segmentation. In *IEEE/CVF Inter. Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 2097–2104. IEEE Computer Society, 2011. doi: 10.1109/CVPR.2011.5995323. URL https://doi.org/10.1109/CVPR.2011.5995323.

Ahmad Humayun, Fuxin Li, and James M. Rehg. The middle child problem: Revisiting parametric min-cut and seeds for object proposals. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

David Stutz, Alexander Hermans, and Bastian Leibe. Superpixels: An evaluation of the state-of-the-art. *Comput. Vis. Image Underst.*, 166:1–27, 2018. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2017.03.007. URL https://www.sciencedirect.com/science/article/pii/S1077314217300589.

Reinhard Klette and Azriel Rosenfeld. The Morgan Kaufmann Series in Computer Graphics. Morgan Kaufmann, 2004. ISBN 978-1-55860-861-0. doi: https://doi.org/10.1016/B978-1-55860-861-0.50027-4. URL https://www.sciencedirect.com/science/article/pii/B9781558608610500274.

Rex Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L. Hamilton, and Jure Leskovec. Hierarchical Graph Representation Learning with Differentiable Pooling, February 2019. URL http://arxiv.org/abs/1806.08804. arXiv:1806.08804 [cs, stat].